Assignment 3

CMPUT 328

Fall 2024

[**Total Weight:** 10%]

1 Classification with Vision Transformer (ViT)

[5% **of the total weight**]

For Part 1, you are required to implement a Vision Transformer (ViT) and perform classification tasks on the CIFAR-10 dataset. The file *vit_submission.py* contains incomplete code for the ViT model and the training function. Marks are allocated to different parts of the incomplete implementation, and your task is to complete the missing code and train the ViT model on the CIFAR-10 train dataset for submission.

You are given `vit_submission.py` and `vit_main.py`. Please do not make any changes to `vit_main.py` as you will not be submitting this file. `vit_submission.py` provides the template code to start your implementation with specific requirements in the comment sections. In addition to completing the code for ViT training, there are a few additional requirements:

- During training, at least 3 data augmentation techniques must be applied.
- During training, a learning rate scheduler must be used.

Failure to meet any of the above requirements will result in **mark deductions** for the respective parts.

1.1 Grading

Please note that your code will not be debugged during grading, and there is **no runtime penalty for this part**. However, any submission that **fails to run** or achieves an accuracy **below the specified accuracy threshold** will receive **no marks**. More details on the marking criteria for Part 1 are provided below:

- **Code (**2.5% **out of** 5%**):** Your code will be assessed for correctness. All missing components in *vit_submission.py* must be completed to receive full marks for this part. Partial marks may be assigned based on the correctness of the implementation.
- **Accuracy (**2.5% **out of** 5%**):** Your ViT model must achieve a minimum test accuracy of 65% on the CIFAR-10 test dataset. Marks will not be scaled linearly, and any submission with a test accuracy below 65% will receive no marks for this part.

2 Image Captioning with Vision Transformer and GPT-2

[5% **of the total weight**]

For Part 2, you are required to implement an image captioning model using the Huggingface Transformer library. Specifically, you will build a Sequence-to-Sequence (seq2seq) model that generates a text caption given an input image. This model will utilize a pretrained ViT (i.e. Google's ViT-Base) as the vision encoder and OpenAI's GPT-2 as the text decoder. The model is to be trained on the Flickr8k dataset. For your convenience, a download link for the Flickr8k dataset with train and validation splits is provided via a Google Drive link.

You are given `cap_main.py` and `cap_submission.py`. Please do not make any changes to `cap_main.py` as you will not be submitting this file. `cap_submission.py` has some template code provided with hints to help you get started. You are free to modify the functions/classes and add your own code, but your submission must return and save a trained image captioning model locally.

2.1 Grading

Please keep in mind that your code will not be debugged during grading. Again, there is no **runtime** penalty for this part. Unlike Part 1, there will be no partial marks for Part 2. Any submission that does not exceed the specified **BLEU score** threshold will receive **no marks**. Detailed marking criteria are as follows:

- **BLEU (5% out of 5%):** Your best performing model will be evaluated over a hidden test set. Any submission with a BLEU score $< 0.07\%$ on the hidden test set will receive no marks for this part. Marks will not be scaled linearly. You should assess the performance of your model using the given validation set before submission, which should reflect the performance of your model on the hidden test set.

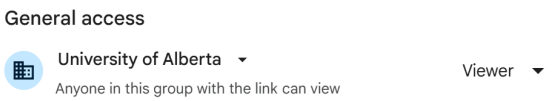
3 Additional Information

3.1 Submission Guidelines

Submit a compressed zipfile as `Assignment3-`$\{YourCCID\}.zip$ containing four files: a) your code implementation for Part 1; b) your code implementation for Part 2; c) your trained ViT model for Part 1; d) your trained image captioning model for Part 2. The zipfile structure should look like:

```
Assignment3-{YourCCID}.zip
├── vit_submission.py
├── vit-cifar10-{YourCCID}.pt
├── cap_submission.py
└── cap-vlm-{YourCCID}.pt
```

Additionally, given that the size of the trained models for this assignment are large. You will need to upload your zipfile onto a Google Drive, and submit the link to the eClass Assignment 3 submission page. Please ensure that you have modified the general access of your submission to "University of Alberta", so that the TAs can access and download your submission for grading.



3.2 Collaboration Policy

This must be your own work. Do not share or look at the code of other students (whether they are inside or outside the class). You can talk to others in the class about solution ideas (but detailed enough that you are verbally sharing, hearing or seeing the code). You must cite **online resources that were referred to and to** whom you talked with, in the comments of your programs.

The usage of ChatGPT is allowed, but not recommended. Additionally, if we have reason to believe that they do not understand the solution that they have submitted, we reserve the right to evaluate any student's submission further through a viva.